FULL-HD HEVC-ENCODED VIDEO QUALITY ASSESSMENT DATABASE

Glenn van Wallendael, Nicolas Staelens, Enrico Masala, Marcus Barkowsky

FULL-HD HEVC-ENCODED VIDEO QUALITY ASSESSMENT DATABASE

*Glenn Van Wallendael, Nicolas Staelens**
Ghent University - iMinds
Ghent, Belgium

Enrico Masala
Politecnico di Torino
Torino, Italy

Marcus Barkowsky
University of Nantes
Nantes, France

ABSTRACT

This work presents a large dataset of High Efficiency Video Coding (HEVC) sequences aimed at creating a reference for researchers involved in video quality research. Ten sequences with different characteristics have been encoded with different parameters, resulting in 59520 videos which have been processed to provide objective quality measurements at frame-level granularity. The ultimate objective is to make this database, which is already publicly available, a reference for researchers involved in designing hybrid video quality measurements, so that it can be used for both investigation and reproduction of results. Besides describing the context and database content, this paper also provides a glimpse on the possible uses of the already available data, as well as the two directions towards which the project is being expanded: incorporating new objective measurements and evaluating quality for corrupted videos with realistic loss traces.

1. INTRODUCTION

The accuracy of objective video quality measurement strongly depends on the suitability of the chosen artifact detection or video quality measurement algorithms and on the training that was performed to optimize prediction performance. A huge number of publications on measurement algorithms exist, a recent overview can be found in [1]. When fusion and training, performance verification, and independent validation of accuracy are targeted, the availability of degraded video databases with ground truth quality scores becomes crucial. Recently a large collection of database links has been made available by Qualinet COST IC 1003 [2]. However, the parameter space of video resolutions, content types, encoding parameters, frame rates, packet loss rates and types, etc. cannot be sufficiently covered by isolated, subjectively assessed video databases.

Therefore, the Joint Effort Group - Hybrid of the Video Quality Experts Group (VQEG-JEG, www.vqeg.org) started an effort towards the creation of a huge database of processed videos. Targeting several million video sequences, reproducibility gains paramount importance for identical reproduction of individual video sequences anywhere on the world and at any time in the future. Fig. 1 shows a flow chart of the processing chain. The left column lists the information that will be stored partially in file space or in relational databases. The middle column contains the processing steps that shall be automated and integrated in a virtualized environment to allow their execution in the near and distant future. On the right side, scientific impact is partially extracted.

In this paper, a first version of the freely available HEVC video bitstream database[1] will be documented, 59520 sequences that were encoded in HEVC format originating from only 10 undistorted source videos, requiring about 313165 computing hours. As the ITU standard guarantees the decodability of these encoded sequences, they can be stored in compressed format. These sequences were evaluated with several Full-Reference measurements and an analysis of their agreement will be presented providing a first view on the usefulness of such a large database. When packet losses occur in the transmission system, the standard does not recommend a particular processing, therefore reproducibility of decoder solutions becomes an issue as detecting erroneous conditions requires either access to the network layer or plausibility tests on entropy decoded information. A possible solution using a modified version of the reference software is proposed before concluding the paper.

2. DATABASE DESCRIPTION

In this section, a description of the current state of the database will be given starting with the selected source videos. Next, the parameters used for compressing these sequences using the HEVC compression standard will be given followed by the full-reference measurements calculated on these processed sequences.

2.1. Source Reference Circuits (SRC)

From the larger VQEG JEG database, ten sequences have been selected similar to the ones used for earlier H.264/AVC

*The research activities described in this paper were partially funded by Ghent University, iMinds, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research Flanders (FWO-Flanders), and the European Union. Some aspects of this work were carried out using the STEVIN Supercomputer Infrastructure at Ghent University.

[1] ftp://ftp.ivc.polytech.univ-nantes.fr/VQEG/JEG/HYBRID/hevc_database

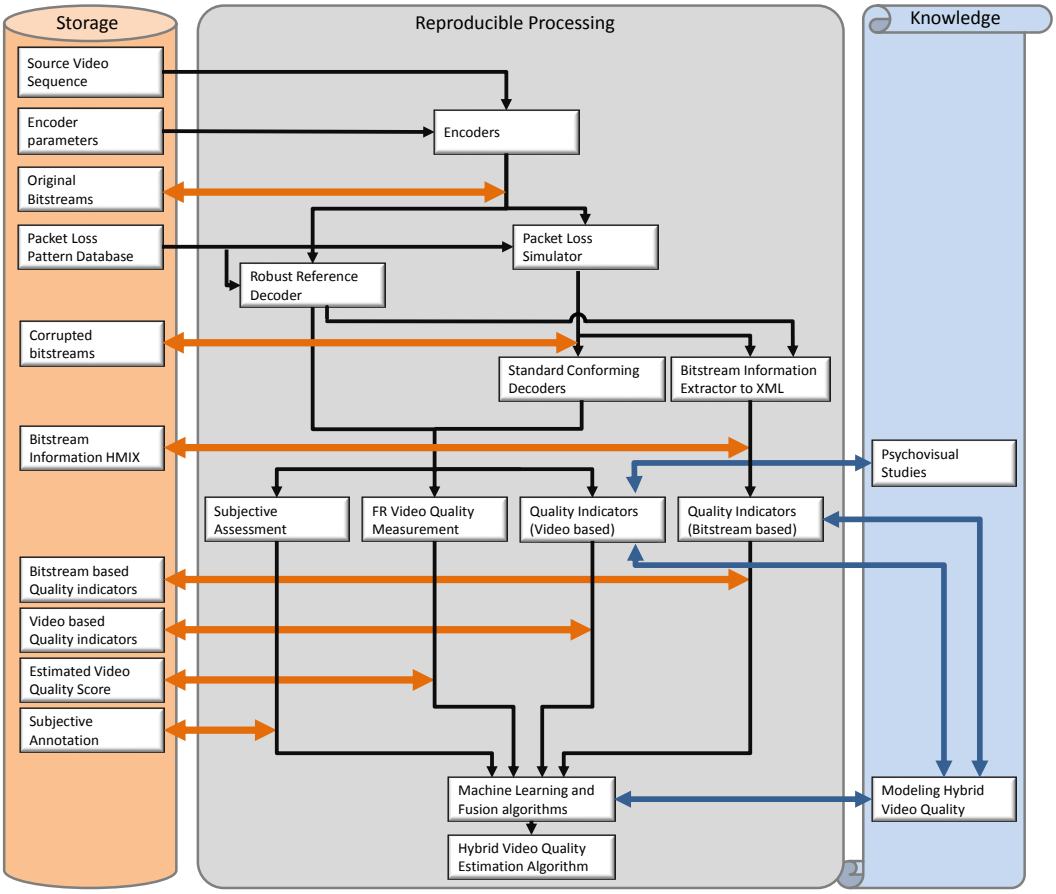


Fig. 1. Processing steps for a large database creation towards development of a reliable Hybrid Model.

based research [3]. Selecting the same source sequences enables possible future comparisons between both compression standards. All sequences are 10 seconds long, play back at 25 frames per second and have an HD (1920x1080) resolution. These sequences were originally selected because they form a diverse set of content types. As illustrated in Fig. 2, the sequences consist of sports sequences, professionally shot sequences, amateur videos, animated content, and so on.

2.2. HEVC Compression Characteristics

Compression of the different sequences is performed using the High Efficiency Video Compression (HEVC) standard. A diverse set of compression parameters is selected in order to cover a broad range of application scenarios (see Table 1).

To start, three different restrictions on bitrate variation have been applied with the most flexible being Variable Bitrate Compression (VBR) using a fixed Quantization Parameter (QP). With this fixed QP configuration, bitrate depends on the complexity of motion and texture within the video sequence. This bitrate variation is mainly applied in scenarios where distribution latencies are less critical like in a broadcasting TV environment or Over-The-Top (OTT) TV service. Less bitrate variation can be imposed by using Constant Bitrate Compression (CBR) on a frame by frame granularity or more strictly on a block by block basis (Coded Tree Unit (CTU) level in HEVC terms). Using less variation, lower latency transmission can be accomplished as needed in a video conferencing application.

Additionally, latency behavior is also influenced by the structure of the Group Of Pictures (GOP). Having a larger GOP-size than one introduces a buffer in the encoder to enable this more efficient frame compression order. For example, with a GOP-size of four, three frames need to get buffered in order to encode the fourth frame first. So, different application scenarios can be covered depending on the GOP-size used.

Finally, an aspect influencing both latency and error resilience is slice size. Using slice sizes of 1500 bytes enables the encoder to send out the frame as soon as it has an Ethernet packet of data encoded. Certainly in low latency scenarios, this can play an important role. The main purpose of slices however is error resilience against packet losses. Therefore, different slice sizes are examined.

For compression of all the sequences using all these different parameters, the HEVC test model (HM) v11.1 has been used. In general, a full matrix of all parameter com-

Fig. 2. Source sequences

Table 1. HEVC compression parameters.

VBR: QP	26, 32, 38, 46
CBR: frame level	0.5, 1, 2, 4, 8, 16 Mbps
CBR: CTU level	0.5, 1, 2, 4, 8, 16 Mbps
Random access	Closed-GOP intra refresh (IDR), Open-GOP intra refresh (CRA)
Intra period	8, 16, 32, 64
Resolution	1920x1080 1280x720 960x544
Slices	Count: 1, 2, 4; Size: 1500 byte
GOP structure	GOP size 1 (IPPPPPPPPP) GOP size 2 (IBPBPBPBP) GOP size 4 (IBBBPBBBP) GOP size 8 (IBBBBBBBP)

binations has been performed on the sequences. However, because of limitations of the software, some combinations could not be performed like a GOP-size of eight using IDR refresh pictures every eight frames. All eligible combinations result in 5952 compression scenarios totalling 59520 video streams in the database.

2.3. Full-Reference (FR) Quality measurements

The quality of all compressed sequences has been evaluated by means of three Full-Reference quality measurements widely used in the literature: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [4] and Visual Information Fidelity (VIF) [5]. Results have been computed using the freely available VQMT tool [6].

3. ANALYSIS OF EVALUATED MEASUREMENTS

One of the first goals of the JEG-Hybrid database creation is the analysis of available objective quality measurements available on the database to spot the presence of some unusual behavior. These anomalies could then be investigated in more details to get some clues about how to design effective quality measurements and in order to understand whether the discrepancies stem from inaccuracies of the algorithms or unveil a known or even unknown property of the Human Visual System (HVS).

Table 2. Correlation between different measurements averaged over the entire sequence.

		Pearson	Spearman
PSNR	SSIM	0.52	0.77
VIF	SSIM	0.93	0.99
PSNR	VIF	0.61	0.81

Table 3. Correlation between different measurements averaged over the entire sequence excluding src09.

		Pearson	Spearman
PSNR	SSIM	0.84	0.97
VIF	SSIM	0.93	0.99
PSNR	VIF	0.94	0.97

For this purpose, first, correlation between the different measurements is evaluated. Pearson and Spearman correlation results between the different measurements can be found in Table 2. In this table, exceptionally low correlation can be observed when correlation with PSNR is computed.

PSNR results as calculated by [6] might be misleading since its value grows to infinity when two frames are identical. This happens in src09 for some frames due to the presence of transitions using black frames which are coded perfectly with respect to the original. By removing src09 from these comparisons, the expected higher correlation can be observed in Table 3.

The original PSNR values are all available in the database, including the case where the value is infinite. This availability allows to reproduce the results of this research. For the following analysis, the values in the database are clipped at 54.15 dB on a frame-by-frame basis. This appears to be a reasonable choice if the PSNR is meant to indicate the end-to-end quality from the analog source until playback. In fact, due to the use of quantized values (8 bit luminance for this database), it is reasonable to assume that there is an unavoidable quantization error, whose average can be considered equal to 0.5. The corresponding PSNR value for this noise level is $10 \log_{10} \frac{255^2}{0.5^2} = 54.15$ corresponding to an MSE equal to 0.25.

Additionally, scatter plots revealing the relationship between the different measurements are given in Fig. 3. These

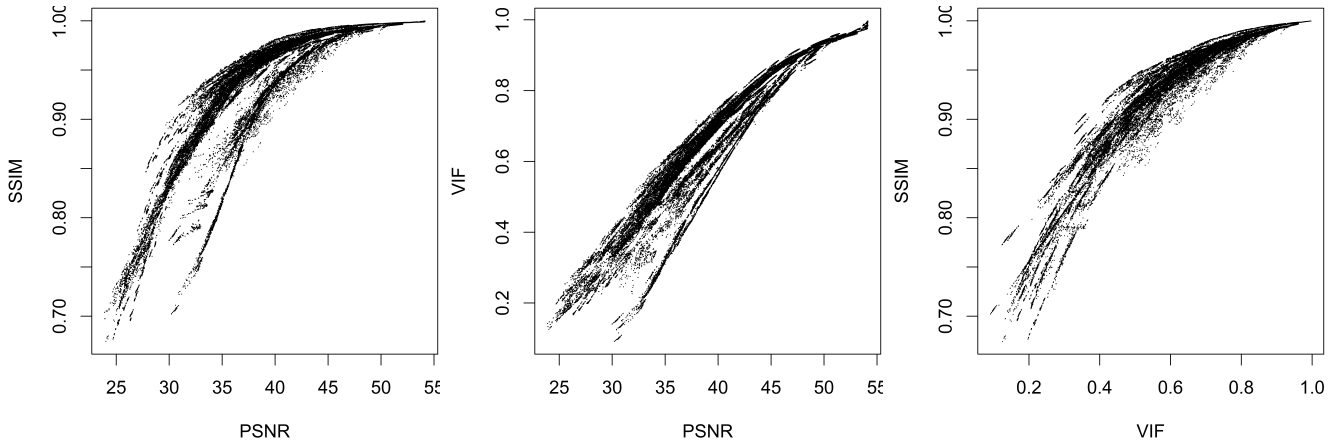


Fig. 3. Comparison between the different evaluated full-reference measurements.

plots indicate that in the high quality range, most algorithms agree about the quality. At the lower quality end, disagreement starts to appear. Therefore, another possibility is to compare all the video sequences in pairs to identify cases in which the algorithms do not agree about which is the sequence with the highest quality. The underlying idea is that if at least one of these measurements does not agree with the indication of the others, this could deserve further investigation. However, the cases in which variations are very limited around the equivalence between the pairs should be treated cautiously since minor variations may be due to tiny, potentially unnoticeable, modifications in the characteristics of the video.

When considering all pairs of sequences in the database (i.e., 1,771,285,440 pairs), in about 10.5% of the cases the three algorithms do not agree on which sequence has higher quality. In particular, disagreement is due to PSNR for about 55% of the cases, SSIM for about 30% and VIF for about 15%. Please note that no measurement performance conclusion can be drawn, it can only be stated that VIF and SSIM agree more often on the ordering. As the algorithmic processing of PSNR, SSIM, and VIF is sufficiently different, interesting modeling conclusions for the HVS arise. Table 4 reports detailed results for each sequence, when only pairs belonging to the same sequence are considered. First, the share of pairs with disagreement is much lower, suggesting that algorithms are heavily influenced by the content. If the analysis is performed per content, the disagreement is reduced. Moreover, reasons of disagreement strongly vary depending on the sequence, as it can be surmised from the columns of Table 4.

To better quantify the algorithms which disagree for each sequence, a normalized difference between all the considered measurements is introduced. For each algorithm, the results are linearly rescaled in the interval [0..1]. Then, the individual differences of all the measurements for a sequence pair are combined in a single normalized difference

Table 4. Reasons of disagreement among quality measurements for each sequence.

Sequence	Pairs with disagreement	Due to PSNR	Due to SSIM	Due to VIF
src01	3.32%	14.47%	60.72%	24.80%
src02	2.64%	40.74%	45.70%	13.56%
src03	6.27%	61.97%	9.30%	28.73%
src04	4.55%	51.17%	11.76%	37.06%
src05	3.30%	37.89%	18.16%	43.95%
src06	4.99%	28.92%	13.84%	57.24%
src07	6.17%	69.45%	7.41%	23.14%
src08	3.93%	24.58%	59.33%	16.09%
src09	7.65%	20.89%	53.62%	25.49%
src10	3.81%	39.76%	12.55%	47.70%

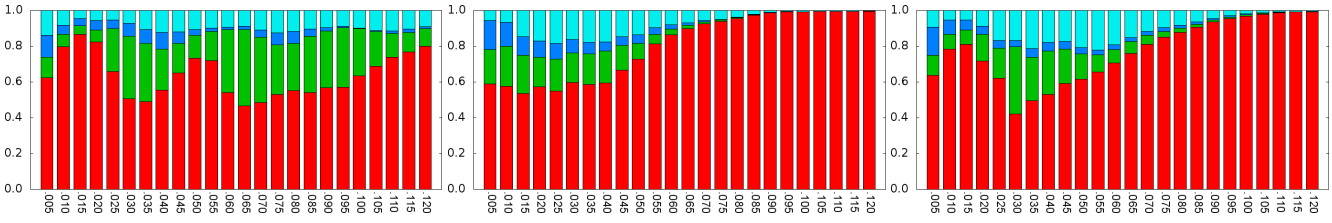
$\hat{d}$ by using the Euclidean distance:

$$\hat{d} = \sqrt{\Delta\widehat{PSNR}^2 + \Delta\widehat{SSIM}^2 + \Delta\widehat{VIF}^2} \qquad (1)$$

so that it is possible to plot all the data in one dimension using a histogram. Figure 4 presents three sample histograms, for src03, src05, src10, showing the reason of disagreement as a function of the normalized difference. Note that the behavior strongly depends on the sequence. For the other sequences the trend is similar. Please note that this is just a sample of the possible results that can be extracted from the database with simple post-processing, showing that there is a large potential in analyzing these data, especially as a function of the coding parameters, to better understand the behavior of the quality algorithms in different cases.

4. FUTURE WORK AND OPEN RESEARCH QUESTIONS

This section covers the parts, already outlined in the introduction, that are still in the early stage of development, giving a glimpse of the near-future activities and the anticipated open research questions. In particular, this section covers the "full-reference video quality measurement", the "packet loss simulator" and the "robust reference decoder".



(src03) (src05) (src10)

Fig. 4. Reason of disagreement (expressed as a ratio over the total pairs) between the various algorithms as a function of the normalized difference for some sequences, shown in brackets (red: agreement, green: due to PSNR, blue: due to SSIM, light blue: due to VIF).

4.1. Additional Full-Reference Quality measurements

Two new algorithms are currently being introduced. The Video Quality Metric (VQM) [7] is a well-established method of objectively measuring video quality and the proposed models have also been adopted as ANSI and ITU standards. This value is already available for the whole set of sequences without losses, also including the 7 intermediate indicators that the model computes for every frame.

The Perceptual Video Quality Measure (PVQM) has been proposed in [8] as a relatively simple algorithm to estimate video quality on the basis of three main features: edginess of the luminance, normalized color error and temporal decorrelation. Also in this case indicators for each frame are available. The possibility to access several indicators at the frame level is expected to strongly contribute to the development of the hybrid video quality estimation model.

4.2. Packet Loss Simulation

The quality of multimedia communication applications might be severely affected by packet losses, hence it is important to investigate how different loss patterns influence the overall quality.

To make experiments reproducible, packet loss models or loss traces should be used. While models might be more interesting from a theoretical point of view to build a complete set of possible combinations of the model parameters, the latter present the advantage to be very realistic.

To foster the idea of reproducible research as much as possible, traces or models should be publicly available. Currently, in the JEG-Hybrid effort, a set of traces has been considered, derived from a public dataset [9]. These traces represent synthetic RTP traffic from a well-connected site (e.g. university) to residential users via DSL or cable connections. Different tests at several transmission speeds are available (1, 2, 5 Mbit/s). We believe that traces of interest may present a packet loss rate (PLR) ranging from 0.005 to about 0.03. The former value roughly corresponds to a single packet loss in a 10-second video sequence at 25 frames per second (fps) with 1 packet per frame. The latter value has been shown to correspond to the lowest quality that can be accepted by the final user of the videocommunication without the insertion of additional robustness techniques.

Regardless of the PLR value, when packet losses are present a robust video decoding software is needed, i.e., the software should be able to handle corrupted bitstreams without crashing and by resynchronizing its internal state with the uncorrupted part of the bitstream as soon as possible. Unfortunately, video decoding software often crash when processing compressed bitstream data corrupted due to missing parts, especially if the amount of lost data is large or affect consecutive elements. Making the software robust to any loss pattern typically requires complex modifications to the processing only available in commercial products and thus inhibiting reproducible research.

Note also that, due to the size of the database considered in this paper (59520 compressed video sequences, requiring 295 GB storage space), storing and distributing the corrupted video sequences resulting from the application of just a small number of packet loss traces does not appear to be a viable solution. A software that can be run locally by any interested researcher should be made available so that the uncompressed distorted video sequences can be reproduced locally, as necessity arises. Also, the availability of the source code could be an important asset for research purposes, allowing to investigate the detailed behavior of the decoder in the cases deemed most interesting. For these reasons, the robust decoding technique described in the next section has been proposed.

4.3. Robust HEVC Decoder

The key idea that allows to transform a generic publicly available decoder (such as the HM test model [10]) into a robust one without the need for an extensive code revision is to avoid affecting the internal state of the decoder, apart from the content of the decoded picture buffer (DPB), when a loss is encountered. This is accomplished by always processing the original, uncorrupted, bitstream and when a loss event is supposed to happen, a simulation of the concealment technique is performed.

In other words, every time a Network Adaptation Layer Unit (NALU) that is supposed to be lost is encountered, the content of the DPB is modified on-the-fly to apply the concealment technique to the areas of the picture that should be affected by the losses. If a simple copy conceament technique, as in [11], is used, the content of the picture in the

buffer is overwritten with the content of the corresponding area of another previous picture which has already been decoded. After this modification, which is executed each time a new NALU is supposed to be lost, the decoder resumes its normal operations so that its internal state is not disrupted avoiding crashes due to unhandled situations in the code.

The main advantage of such an approach is: 1) avoiding software crashes due to data loss (any loss pattern can be handled); 2) the possibility to implement different concealment techniques; 3) the possibility to realistically simulate the reconstructed video by a hypothetical decoder that operates on a real corrupted bitstream by making small modifications to the official HEVC HM test model software [10]. Note that in most cases the reconstructed video is exactly the same as the one that would be produced by a decoder operating on a real corrupted bitstream using the same concealment technique. However, in very few cases there might be a slight misalignment, for instance if some coding modes that require the availability of data from previously decoded frames incorrectly assume that the data is available when it would not (e.g., some particular cases of motion vector predictors computation). Nevertheless, the type of artifacts resulting from this approach are very similar, when not exactly the same, to the ones of a decoder operating on a real corrupted bitstream.

The possibility to freely distribute such a software (allowing any interested party to inspect and use it) overcomes this slight limitation and it is perfectly in line with the aims of the JEG-Hybrid project. The software is available[2] as a modification to the HM test model software [10]. Since it is not feasible to directly distribute the corrupted video sequences, the project makes the measurements of all the video sequences as corrupted by losses publicly available in the database[3], so that they can be used by any interested researcher. Moreover, the strong focus on reproducible processing will allow anybody to recreate all the processed data locally as needed.

5. CONCLUSIONS

This work presented a large dataset of HEVC-coded video sequences that can be used as a reference for researchers involved in designing hybrid video quality measurement algorithms. Such a database can be used both for designing new algorithms and as a reference in other research efforts to allow result validation. The main advantage of having this publicly available database, in which all results are reproducible, is to avoid researchers to go through the expensive process of creating such a large reference database that requires months of efforts, in terms of both human resources and computational time, so that they can readily focus on the research efforts more related to video quality. The database is currently being expanded in two main directions: the inclusion of more sophisticated objective quality measurements and the consideration of videos corrupted due to transmission over packet networks. Finally, every researcher working in the field is welcome to join the effort by, for instance, contributing new quality measurements or software for processing and analyzing the data.

[2]http://media.polito.it/jeg

[3]http://media.polito.it/downloads/jeg/with_losses

6. REFERENCES

[1] S. Chikkerur, V. Sundaram, M. Reisslein, and L.J. Karam, "Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison," *IEEE Transactions on Broadcasting*, vol. 57, pp. 165–182, Jun. 2011.

[2] K. Fliegel, C. Timmerer, (eds.), "WG4 Databases White Paper v1.5: QUALINET Multimedia Database enabling QoE Evaluations and Benchmarking," http://dbq-wiki.multimediatech.cz, Mar. 2013.

[3] M. Barkowsky, N. Staelens et al., "Subjective experiment dataset for joint development of hybrid video quality measurement algorithms," in *Workshop on Quality of Experience for Multimedia Content Sharing*, Berlin, Allemagne, Jul. 2012, pp. 1–4.

[4] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[5] H.R. Sheikh and A.C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[6] P. Hanhart and R. Hahling, "Video quality measurement tool (VQMT)," http://mmspg.epfl.ch/vqmt, Sept. 2013.

[7] M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transaction on Broadcasting*, vol. 50, no. 3, pp. 312–322, Sept. 2004.

[8] A.P. Hekstra and J.G. Beerends et al., "PVQM – a perceptual video quality measure," *Signal Processing: Image Communication*, vol. 17, no. 10, pp. 781–798, Nov. 2002.

[9] M. Ellis, C. Perkins, and D.P. Pezaros, "End-to-end and network-internal measurements on real-time traffic to residential users," in *Proc. of ACM Multimedia Systems*, San Jose, CA, USA, Feb. 2011, pp. 111–116.

[10] K. McCann, B. Bross, W.-J. Han, I.-K. Kim, K. Sugimoto, and G. J. Sullivan, "High Efficiency Video Coding (HEVC) Test Model 12 (HM 12) Encoder Description v. 12.1 Doc. JCTVC-N1002," Nov. 2013.

[11] S.K. Bandyopadhyay, Z. Wu, P. Pandit, and J.M. Boyce, "An error concealment scheme for entire frame losses for H.264/AVC," in *Proc. of IEEE Sarnoff Symposium*, Princeton, NJ, Mar. 2006.